Constructive Alignment, Engagement and Exam Performance: It's (still) ability that matters.

Mary R Hedges[1] and Gail A Pacheco[2]

[1] *Centre for Longitudinal Research, University of Auckland, New Zealand*
[2] *Department of Economics, Auckland University of Technology, Auckland, New Zealand*

Abstract

The use of online quizzes has become more popular in introductory economics courses in recent years however their efficacy in improving student engagement and performance has seldom been evaluated. This paper attempts to undertake evaluation of that efficacy by combining individual engagement and performance data with demographic data for a cohort of students enrolled in a first year economics course. It was found that demographic characteristics only impacted indirectly through the level of engagement by the students. The use of bivariate probit modelling to further examine the relationship between quiz efficacy (as an indicator of a student's level of productive engagement) and exam performance was a novel extension to this type of analysis. These results indicated that irrespective of whether the student has above or below average engagement in their course, the marginal effects of all other observed individual characteristics are similar in terms of their impact on final exam mark. In particular only the student's underlying ability stood out as having a significant impact on exam performance, once we had conditioned our results on the engagement variables.

Keywords: Threshold concepts, student engagement, ability, bivariate probit, constructive alignment

JEL codes: A22

Corresponding author: Mary Hedges, Centre for Longitudinal Research, The University of Auckland, Private Bag 92-019, Auckland 1142, New Zealand. Email: m.hedges@auckland.ac.nz

1. Introduction

In 2009 the Bachelor of Commerce (BCom) at the University of Auckland was subject to a major curriculum review. The outcome of this review was the introduction of two new, integrated first year business papers into the core of the degree (University of Auckland Business School, 2009). One of the costs of this change was that economics would be reduced from two compulsory papers in the first year to only one (hereafter referred to as 191). The second implication of this wider curriculum review was that this new first year core paper would be common to all three degrees offered in the Business School: Bachelors of Commerce, Property and Business Information Management. These changes were to be introduced for the 2011 academic year starting with semester 1 in March that year.

Prior to 2011, students did a full semester of microeconomics (101) and macroeconomics (111), the former being a prerequisite for the latter. In contrast 191 was developed as a mixed micro- and macroeconomics paper[1]. For students not majoring in economics this would probably be the only economics paper that they would take in their business degree. Therefore it was imperative that the course would teach students the basic principles of economics and how they could be applied to a wide range of real world situations. At the same time the new course needed to meet external accreditation requirements including NZICA (New Zealand Institute of Chartered Accountants) and international business school accrediting programmes (AACSB and Equis)[2]. For students who do major in economics the course had to prepare them to progress to higher level economics papers without the student having to take an additional course. Having to take an additional course would make economics a less attractive major option compared to all other majors available which could impact on enrolments.

Whereas many course re-developments begin by considering the course content that needs to be added or removed, our course re-design was treated as a green-fields process that started with consultation with industry, faculty and accrediting agencies. Through these consultations a list of learning outcomes was developed. This process not only considered the topics and technical skills that were required but also how the students would be able to apply that content to a range of real world situations. That is, the focus was strongly on capability building rather than just content acquisition. It was only at this stage, mapping the new outcomes and content against the existing ones took place.

The resulting course was one in which the ideas and concepts of constructive alignment (Biggs, 2003) were applied. Consequently, the content, teaching methods, application of theory and assessment items were all aligned with the main purpose of meeting the core learning objectives, i.e. developing the students' application of economic concepts and critical thinking skills. Furthermore, the assessment programme became a key teaching tool instead of acting primarily as a measurement tool (Rowntree, 1987). That is, the goal was to use the assessment strategically to encourage the type of engagement and integrative learning that we were seeking (Gibbs, 1999).

Understanding the pressure of assessment in single semester (12 week) courses there was a sub-text of keeping students on task and their economics course at the top of their priority list without overloading them. Following ideas of constructive alignment we wanted students to:

- get early feedback on their understanding, and reflect on this feedback
- be able to apply theory to real world problems (not seen before)

[1] See Hedges (2012) for greater details on the curriculum re-design, background, solutions and implementation issues. This paper describes the pedagogy theory of constructive alignment as underlying the re-development of 191 to meet the objectives of the new degree programmes.

[2] Association to Advance Collegiate Schools of Business and European Quality Improvement System respectively.

- to remember what they had learned

The purpose of this paper is to statistically analyse whether this constructively aligned approach had a positive outcome on student results, with respect to their final exam mark. We also apply a bivariate probit model to identify whether the impacts of demographic and other individual characteristics on exam performance differ depending on whether the student has above or below average engagement in their economics paper (via a variable that interacted time spent and mark gained on weekly online quizzes). The remainder of this paper is organized as follows: Section 2 provides a brief literature review defining and linking the concepts of constructive alignment and student engagement with exam performance. Section 3 outlines the data used, and details the specific assessment structure developed for this economics course. Section 4 describes the methodologies employed. Finally, Section 5 details the results and concludes.

2. Literature review

The development of 191 was based on the principles of constructive alignment (Biggs, 1999 2003). That the course (and wider programme) should seek to engage the students in a range of learning activities that produced compatibility between curriculum, teaching methods and assessment procedures (Taylor, 2002). However, having established what the topics and outcomes required were the challenge then became to constructively align the curriculum to achieve these aims and objectives subject to a number of significant constraints. Among these constraints were:

- expected large numbers (up to 1500 students per semester once the new structure was fully introduced[3]),
- resourcing limits
    - contact time limited to 3 hours of large lectures (up to 600 students)
    - potential for a one hour tutorial per week (class sizes up to 30 students)
    - marking assessment constraints including:
        - assessments in an essay format would be difficult for many markers to deal with effectively, consistently and in an appropriate time frame (markers are predominantly post-graduate students within the economics department)
        - timeliness of feedback given large size of class.
        - early feedback to ensure concepts are understood
- lecturer constraints to cover the number of streams and maintain consistency between streams
- core theory coverage to enable progression to higher level economics papers.

It quickly became obvious that the only way to achieve our aims, within the bounds of the constraints noted above, was to take a 'less-is-more' approach. This was based on the pedagogy of threshold concepts (Davies, 2003; Davies & Mangan, 2005, 2008; Land, Cousin, Meyer, & Davies, 2005; Land & Meyer, 2006; Meyer & Land, 2005), where there is a large body of literature (within the economics discipline) to draw on. Central to many of these studies is the 'Embedding Threshold Concepts' project funded jointly by the HEFCE, the Institute for Education Policy Research, Staffordshire University and Enhancing Teaching-Learning Environments in Undergraduate Courses projects in the UK (Staffordshire University, 2008). This project had produced a number of resources to aid curriculum design based on these ideas. Also, fortunately, economics has a very standard global body of such concepts that can be

[3] Transitional arrangements were agreed to for the first two years of the new degree structure where students could elect the old micro-macro option or the new single course option.

applied at the local level, with a New Zealand perspective.

Considering the curriculum in terms of threshold concepts meant actively considering what these concepts are and how they differ, or remain similar to foundational concepts which are what the standard 101/111 courses tended to be based on. Threshold concepts by definition are considered to be a 'portal' or 'conceptual gateway' (Meyer & Land, 2003) to the discipline. Meyer & Land (2003) define the characterisation of these conceptual gateways as being:

- bounded
- integrative
- transformative
- potentially troublesome
- probably irreversible

They have been placed in this specific order due to the path dependency required in developing some of these ideas (Hedges, 2012).

In contrast, fundamental concepts are considered 'building blocks' or 'core ideas' that are considered essential to progress learning in the discipline. These are often used in curriculum design and were certainly the basis of 101/111. Their use may result in a 'theory first' approach that sees complex ideas simplified and risk 'rote' learning of methods that lose of sight of deeper ideas. The focus on foundational concepts often divorce understanding from experience or at best relates theory to experience as an example (Davies & Mangan, 2005). In contrast threshold concepts keep the application as central to the concepts taught. Foundational concepts are very useful in distinguishing one type of economics from another but less useful in distinguishing economic thought from other disciplines (Davies, 2003). Clearly in focussing on threshold concepts in the re-design of 191 it was a fundamental shift from the basis of 101/111 development.

Key components of threshold concepts are that they are integrative and transformative. However, they can only meet these objectives if the student is actively engaged in their acquisition. Student engagement is variably defined as when:

> "Students are making a psychological investment in learning. They try hard to learn what school offers. They take pride not simply in earning the formal indicators of success (grades), but in understanding the material and incorporating or internalizing it in their lives" (Newmann, 1992) and
>
> "they are involved in their work, persist despite challenges and obstacles, and take visible delight in accomplishing their work." (Schlechty, 1994) and
>
> "Student engagement also refers to a student's willingness, need, desire and compulsion to participate in, and be successful in, the learning process promoting higher level thinking for enduring understanding." (Bomia et al., 1997)

Clearly student engagement overlaps with both student motivation and achievement but is not the same as either. Many measures of student engagement such as the National Survey of Student Engagement (NSSE) or the College Student Experiences Questionnaire (CSEQ) are focussed on the time a student spends on certain types of behaviours that have been observed to be highly correlated with many desirable learning and personal development outcomes (Axelson & Flick, 2010). Unfortunately this type of measurement is focussed more on the students' inputs into their learning experience rather than necessarily their efficacy at converting those inputs into the desired outcomes. Measuring student engagement as time spent on certain learning related behaviours can obscure some of the relationships between the behaviour and the effectiveness of the learning opportunity, that is, the student's efficacy at converting the behaviour into learning. This has become the focus of more recent literature (Galizzi, 2010; Kuh, 2010; Park &

Kerr, 1990) and is a key interest when investigating classroom innovation and researching effective practice. This research seeks to add to that literature by focussing on the efficacy of engagement behaviours, particularly the use of online quizzes. For this reason the terms quiz efficacy and productive engagement are used interchangeably when interpreting the results of this study.

3. Data

We employ data from the first cohort in this newly designed 191 economics course. In addition to the summative assessment results we have very rich data on student quiz participation, including number of attempts on each quiz, time spent on each attempt and marks gained. This course data was then supplemented with matched enrolment data [4] to enable controls for demographic and other individual characteristics of each student such as ability. This process of matching data sets was done by using the student ID number, which was then removed from subsequent data sets, to ensure anonymity under the conditions of the ethics approval. The descriptive statistics and definitions of all key variables are shown in Table 1.

< Insert Table 1 here >

The sample consists of 555 students and is relatively evenly split across gender, with just over 55% of the group being male. Relative to the control group for ethnicity (Pakeha), there are two other groups large enough to enable effective and meaningful analysis; Maori and Pacific Peoples (MaPP), and Asians, who account for approximately 10 and 56 per cent of the sample respectively. Observed underlying ability of the student (regardless of course structure) is collected via their cumulative Grade Point Average (GPA) which indicates a wide range of students in this sample, with GPAs ranging from 0 to 8.75 out of a possible 9[5].

In terms of the assessment indicators, we have information on all three components of assessment (Tutorials, Quizzes and Test) prior to their exam performance. As a general rule, any student who did not complete at least three of the four assessment items was removed from the sample [6]. This has the potential to upwardly bias the final grade distributions, however, comparison of these distributions with those from previous semesters show that they are similar, and well within the Faculty guidelines. It is also important to remember that the objective of the course re-design was not to influence grade distributions *per se*, but to improve critical thinking skills, application and recall of course content.

Assessment Structure

To these ends the assessment structure was developed as an integral part of the entire course. While online quizzes had been used offered as formative options only in the previous first year economics papers, the take-up of them when formative only tended to be very low once assignment pressure started. At the same time, the literature suggests (Cameron, 2010) that even when items carry very low mark weightings, they can act as a great stimulator for student

[4] This analysis uses university level data on individual students that was not collected for the purpose for which it will be used and therefore introduces ethical considerations. Ethics approval was gained through the University of Auckland Human participants Ethics Committee, September 2011, reference 7570.

[5] This cumulative GPA does include the student's grade in 191. For this reasons this GPA will be biased slightly towad their economics grade. This bias will be reduced the greater number of papers that student ha include din the cumulative GPA.

[6] This involved the elimination of just 10 students from the 570 who completed at least a part of any one assessment. Of those ten none of them had sat the test or the final exam and had done only one or two quizzes and/or tutorials. All of these students appear to have made an early decision to not complete the course but did not formally withdraw.

engagement and completion. If the assessment programme was to be used to help teach key skills and capabilities as well as to assess their learning, marks were therefore required to show the priority of all tasks. This need had to be balanced with the unsecured nature of the quizzes and tutorials.

In developing this integrated assessment programme the teaching and learning objectives were explicitly aligned and this structure explained to the students. The structure was also consistent with the explicit capability development goals of each item. Weekly quizzes were designed to encourage mastery of the core ideas and theories. Fortnightly tutorials then assumed this mastery and developed application capability through the analysis of real world problems in a systematic and overt way. The test then expected students to apply this capability to different real world problems they had not seen before. Based on this it would be expected that students that had mastered all three of these components would be much better prepared to succeed in the final exam which was similar in structure to the test.

The weekly quizzes (run through an external platform) were designed to ensure students were understanding the core theory and ideas and able to apply them to simple text book situations (Buckles & Siegfried, 2006). Feedback would be automated and immediate on these quizzes. This structure managed the very real resource constraints faced by large first year papers where up to 1,500 students were enrolled per academic year. The five, fortnightly tutorials were where more complex, real world examples would be addressed, discussed and solutions found using the theory taught and tested through the quizzes. While the preference would have been for weekly tutorials, room and tutor constraints meant this was unrealistic given the size of this course. The mid-term test and final exam would then require students to analyse and apply the material to real world problems that they had not seen in this context before.

The details of the programme settled on was:

- 11 weekly quizzes (10 marks)
    - Best ten counted
    - Unlimited attempts but only the best would count
    - Assesses key theory and basic understandings
    - 10 day window to complete each weeks test
- 5 tutorials (10 marks)
    - Applied theory to real world problems
    - Stepped through process of application but...
    - …then had to attempt another one their own and hand it in for marking and feedback.
- Mid-semester test (30 marks)
- Final exam (50 marks)

Both the mid-semester and final exam would then present students with a mix of recall of key concepts and the application of those concepts to current issues. The weighting between these components was toward the application (70%) with the recall questions being present primarily to build confidence in the time constrained test/exam environment.

4. Methodology:

The first model employed is a linear regression (using ordinary least squares) where exam mark will be the dependent variable and various demographic and assessment variables will be the determinant variables. The exam mark was chosen as the dependent variable in preference to the final course mark because while the exam mark was dependent on a student's engagement and ability to recall and apply the course material it was independent of their performance on the

separate assessment items. In contrast the final mark in the course was made up of the cumulative mark of all assessment items. For both the tutorials and the test the mark achieved was used as the most appropriate metric as this captured both in class participation and efficacy (in the case of the tutorials) and was the only metric available for the test. However, for the quizzes there were a range of potential engagement metrics available (mark, number attempted/completed or time spent which could be total, average or median). Some understanding of the quiz structure is necessary to understand why an interactive variable was chosen as the most appropriate measure for the subsequent analysis.

There were a total of eleven weekly quizzes available for students, each out of 15 marks, however only their top ten quiz marks would count. This built in redundancy was to minimise issues if students had technical or personal problems that would result in them missing a quiz. Rather than having to deal with these issues the option of the extra test meant that they were not penalised for one bad week. There was no time limit on each quiz attempt but each quiz was only available for ten days. For each quiz there was then a quiz bank of approximately 150 questions on that week's material. These questions were a combination of questions taken form the publisher test bank related to the text book and additional questions developed by the teaching team. All questions were checked, and if necessary edited, to ensure they were correct for the New Zealand economy. Each bank of questions was split into 15 pools that matched the topics covered that week. Each randomly generated quiz had one question from each pool. This structure ensured that every quiz generated covered all of the material required for that week. It would not be possible for a student to 'get lucky' and get questions on only one or two of the ideas covered in lectures. In addition to this layer of randomisation, the order of the answers for most of the multi-choice questions would change each time a quiz was generated. Thus, although a student may get a question that they had seen before the order of the possible answers would still be different. Some questions were calculation questions based on algorithms so that the answer would be uniquely different each time the student saw the question.

Based on the complexity of this quiz structure it would be virtually impossible for a student to do sufficient quizzes to memorize all the answers. This was the basis on which the decision was made to allow students an unlimited number of attempts at the quizzes within the ten day timeframe that the quiz was open. It was also discussed in class that it did not matter how they used the quizzes. They could do all the work and study before attempting a quiz and use it to check their learning in which case they may only do each quiz once and get full marks. Alternatively they could do the quizzes multiple times and use them as a tool to learn the material in which case they may spend a lot of time and attempts within the quiz environment in order to get the same full mark result. Obviously some hybrid of these two extremes would also be possible. This flexibility in how students could choose to use the quizzes meant that the most appropriate metric was not obvious.

If mark was chosen this would give an indication of final outcome but not provide any indication as to how it had been achieved. If total time spent doing the quizzes over the eleven weeks (or median or average) was used this would combine and average the two approaches concealing variation in type of engagement. In graphing mark gained against average time spent a slightly u-shaped relationship was found with higher marks being gained by those who spent little time on the quizzes because they have done the work before attempting the quiz and those who spent a lot of time on average because the quizzes were their mode of engagement. To accommodate the variation in the type of engagement a new variable was calculated by interacting the time spent with their final quiz mark in order to provide a measure of the efficacy of their quiz engagement. That is, how effective were the quizzes in enabling the students to develop mastery

of the material? This was the only item that we had sufficient detailed data in order to analyse in this way which is why the quiz is used as a proxy for engagement in the analyses undertaken.

When introducing these variables into the OLS regression each assessment item was introduced alone initially and then removed before the next one was used. The reason for this approach is that under a constructive alignment paradigm it would be expected that no or all individual items would matter but that the cumulative effect of the items would be significant in determining final grade. Although not reported here it was found that in each of these early models the assessment item introduced was significant at the 1 per cent level and the order of magnitude was similar for each. This supports the constructive alignment approach that each item was important though each was possibly capturing a more general engagement factor when introduced in isolation. To attempt to control for this the second approach was to add the assessment items in their correct order, that is, the quiz efficacy followed by the tutorial mark and then the test mark. These results would then demonstrate the cumulative effect of the integrated and developmental approach that underlies this course and separate out the more general engagement form the contribution of each component item.

The second empirical model employed in this study is a bivariate probit, which assumes the data takes the sequential format shown in Figure 1. This paper examines whether there are associations between a range of individual characteristics on quiz efficacy and exam performance. A distinctive feature of this analysis is that our methodology allows us to model the influence of the control variables on both exam performance and quiz efficacy at the same time.

<Insert Figure 1 here >

Figure 1 reveals that 57.3% of the sample performed below average in terms of our constructed quiz efficacy variable. It appears clear that those that performed above average were more likely to also perform above average in their exam (65.4% versus 51.57%). The scenario presented in this figure involves the analysis of two dependent, and sequential variables to model. This permits marginal effects to be obtained where P(Exam mark = 1| A or B), i.e. P(Exam mark=1 | Quiz efficacy=0) and P(Exam mark=1 | Quiz efficacy=1), where one denotes above average, and zero equates to below average. Given these marginal effect estimates from these two conditional probabilities (i.e. comparing route C with route E) it is possible to then identify whether the drivers of exam performance differ depending on whether quiz efficacy=1 or 0. This model therefore allows empirical investigation of whether the determinants of above (below) average exam performance are measurably and statistically different, depending on whether the student undertakes above or below average quiz efficacy.

More formally, let y_{1i} be a latent variable that denotes the probability that a student has above average exam performance, which is determined by a range of explanatory variables, X_{1i}. Also let y_{2i} be a latent variable for the probability the student has above average productive engagement as measured by quiz efficacy, which is also determined by a range of explanatory variables, X_{2i}. The model is represented as follows:

$$y_{1i} = \beta_1 X_{1i} + \varepsilon_{1i} \quad (1)$$

$$y_{2i} = \beta_2 X_{2i} + \varepsilon_{2i} \quad (2)$$

where y_{1i} is observable for all students in the sample and related to the following binary dependent variables, on the basis of the following conditions:

$Exam_i = 1, \text{if } y_{1i} > 0 \qquad Exam_i = 0, \text{if } y_{1i} \leq 0$

and

$quiz_i = 1, \text{if } y_{2i} > 0 \qquad quiz_i = 0, \text{if } y_{2i} \leq 0$

where $Exam_i = 1$ denotes that the student has performed above average in their exam and within their cohort, and $quiz_i = 1$ denotes that the student has above average values for our constructed indicator for quiz efficacy. The errors $(\varepsilon_{1i}, \varepsilon_{2i})$ are assumed to have the standard bivariate normal distribution, with $E(\varepsilon_{1i}) = 0 = E(\varepsilon_{2i})$. The bivariate probit model has full observability as both $Exam_i$ and $quiz_i$ are observed in all four of their possible outcomes (i.e. '$Exam_i = 1, quiz_i = 1$', '$Exam_i = 1, quiz_i = 0$', '$Exam_i = 0\ quiz_i = 1$' and '$Exam_i = 0\ quiz_i = 0$',), which leads to the most efficient estimates (Ashford & Sowden, 1970; Zellner & Lee, 1965).

5. Results:

Table 2 presents the results of three linear regression models. In all specifications, individual characteristics and the type of degree the student is enrolled in are included as determining factors behind exam performance. In model (1), our interactive measure of productive engagement is also included – quiz efficacy. In the subsequent regressions, the model builds to add our other assessment metrics, tutorial mark and then test mark (models (2) and (3) respectively) in an order the reflects the items role in capability development as discussed in sections 3 and 4.

< Insert Table 2 here >

In general, males appear to have performed better in their 191 exam, relative to females & this result is statistically significant at the 5 per cent level in the first two of the regression models. This finding is consistent with evidence provided by Krieg & Uyar (2001), who also find that males outperform females in their analysis within the economics discipline. At first this result may seem strange, because other studies find that females tend to be more engaged (Kuh, 2010), and we would expect that increased engagement would lead to improved exam performance. It is important to note that this link may depend on the nature of the exam. Hickson (2010) found that when exams were in the format of constructed response answers, women did better while males tended to do better when the questions were multi-choice or numerical. This exam was a mix of question types but with an application focus throughout. This may explain the results found and why they were only significant in the early forms of the model.

There appears to be a negative relationship between age and exam performance, albeit statistically insignificant. This effect is consistent with other literature that draws attention to the range of reasons that mature students may not perform as well and tend have a higher drop-pout rate than their younger counter-parts. These include: a lack of preparedness for higher education; changing personal circumstances or interest; financial matters; the impact of undertaking paid work; and dissatisfaction with the course or institution (Duff, 2004). A later extension of this initial study could be to control for full-time or part-time status and see if this relationship holds or is mitigated by study mode.

Both ethnic minorities (MaPP and Asian) fared slightly worse in their 191 exam, relative to the control group of Pakeha and this is statistically significant at the 10 per cent for the Asian group in the fullest specification. Interestingly the coefficient is positive for Maori and Pacific Peoples in the second specification. This specification includes discussion and collaborative learning with their peers within the tutorials which has been linked to cultural learning norms for these ethnic groups pedagogy. The tutorials also provide an greater opportunity for relationships to be developed with tutors that are not always possible in the large lecture formats and are known to support the learning in these groups in particular (Gorinski & Abernethy, 2007). Although these results are consistent with other studies they are generally not significant. Perhaps related to these findings is that domestic students do slightly worse relative to international students in all specifications of the model but this is only significant at the 10 per cent level in this first two specifications. This again has the expected sign based on the literature (Kuh, 2010).

The student's underlying ability (as measured by their cumulative GPA) is a strong and consistent predictor of exam performance, significant at the 1 per cent level in all three specifications. All studies consistently find that underlying ability is the key determinant of future success (Anderson, Benjamin, & Fuss, 1994; Davies & Guest, 2010; Galizzi, 2010). In the upcoming empirical analysis, where a bivariate probit model is employed, this study can help untangle the complex relationship between engagement and exam performance, and investigate the extent to which (if any) higher engagement can compensate for lower ability.

The impact of quiz efficacy is also significant at positive & importantly this result is consistent in all three models, even as other assessment metrics are added to the specification. As the additional items are added the level of significance of the quizzes does decrease. However, that all three capability building assessment items are significant tends to support the achievement of constructive alignment within this course. That quiz efficacy remains significant and is the best measure that captures both time spent and the productivity of that engagement is further motivation for the use of this metric as our measure of productive engagement. Therefore when separating the empirical analysis into two sequential steps in the bivariate probit model it is consistent to use the interacted quiz measure as the first step a student undertakes within their enrolment in 191, and exam performance as the second step. As shown earlier, in Figure 1, it is clear that those that performed above average in quiz efficacy were more likely to also perform above average in their exam. Is it possible to get an idea on the causal pathway of this relationship?

Bivariate probit

In the linear regressions contained within Table 2, quiz efficacy is treated as an independent variable associated with exam performance. While that approach may be valid, a possible further perspective is that quiz efficacy and therefore productive engagement by a student is a necessary prerequisite step prior to the impact of other determinants on exam performance. It is possible that allowing this sequential process between quiz efficacy and exam performance illustrates whether the student's level of engagement exacerbates or diminishes the impact of other influences on the exam result. Consequently, the bivariate probit model presents an empirical investigation into whether the determinants of exam performance (above or below average result) differ depending on the level of engagement a student experiences in their 191 course. Essentially, this involves estimating equation (1) and (2) together (based on the belief they are related), and then extending this to the sequential form detailed in Figure 1, as the marginal effects of equation (1) are estimated conditional on the results of equation (2).

Application of the bivariate probit model obtains results presented in Table 3. The first column corresponds to the determinants of above average quiz efficacy (equation (2)). These results are in line with expectations; males are less engaged relative to their female counterparts; Asians are more engaged relative to Pakeha (Kuh, 2010); and underlying ability and the other assessment measure of tutorial mark have a positive relationship with quiz efficacy. The second column of Table 3 illustrates the determinants of exam performance, comparing routes C and E with D and F on the tree diagram in Figure 1. Apart from the influence of test mark, the only individual characteristic impacting exam performance is the student's underlying ability. Given that the test is of similar format to the exam and is an opportunity to practice that format of assessment this relationship is as expected (Hickson, 2010). The biprobit identifies whether the regressions are related. This is given by the rho (= -0.052), which is statistically significant (p=0.095).

< Insert Table 3 here >

Table 4 presents the estimation of conditional marginal effects of the explanatory variables on exam performance once we've estimated the impact of the explanatory variables on quiz efficacy. This is essentially a comparison of routes C and D on the tree diagram in Figure 1, with routes E and F, i.e. the determinants of exam performance with above and then below average quiz efficacy. There are two key findings in this table. Firstly, the determinants of exam performance are relatively stable, irrespective of whether or not the student experiences above or below average engagement via their online quizzes.

< Insert Table 4 here >

Secondly, is the empirical investigation of whether the determinants of above (below) average exam performance are measurably and statistically different, depending on whether the student undertakes above or below average quiz efficacy. Leaving aside our assessment indicator of test mark[7], it appears that only underlying individual ability, as observed via their cumulative GPA, remains as a determinant of exam performance once we have conditioned our results on the variables that influence engagement. This means that, based on this cohort of students, when two students have the same level of engagement the only thing that will influence their exam mark is their underlying ability (Duff, 2004; Fizel & Johnson, 1986; Shanahan, Foster, & Meyer, 2008).

Conclusion

This paper has attempted to analyse the effectiveness of a constructively aligned course structure that uses assessment explicitly as a learning tool for students. The course was designed to encourage greater student engagement and connection with the course material. The data available allowed us to control for numerous individual characteristics in an attempt to separate out the impact of the student's level of engagement in course as measured by the component assessment items. It was found that all assessment items positively contributed to final exam mark.

The OLS regression analysis indicated that our results were consistent with the literature on the demographic impacts on performance. Our results had the right direction of impact but were often not significant. For example, age and ethnicity were generally not significant in this study. Given that we had data to include engagement measures in the regression it is expected that

[7] Note – the expected learning outcomes of the test and exam are very similar.

these demographic features impacted indirectly through the level of engagement by the students as so was captured in those variables rather than the more direct influence found in other studies that did not have engagement measures available. What was important in terms of this course design was that all of the assessment components were important in determining final exam mark. The course was designed to explicitly build the content knowledge, skills and capabilities required to apply economic theory to real world problems as tested in the exam. These results support this hypothesis.

This use of bivariate probit modelling to examine the relationship between quiz efficacy (as an indicator of a student's level of productive engagement) and exam performance was a novel extension to this type of analysis. The results indicated that irrespective of whether the student has above or below average engagement in their 191 course, the marginal effects of all other observed individual characteristics are similar in terms of their impact on final exam mark. In particular, apart from the influence of their test mark (extremely similar to their exam), only the student's underlying ability stood out as having a significant impact on exam performance, once we had conditioned our results on the engagement variables.

The analysis undertaken in this paper goes some way to understanding the effectiveness of a constructively aligned course in meeting the learning objectives. However, it would be expected that collecting better measures of student engagement, particularly the efficacy of that engagement would further enhance these understandings. The importance of underlying ability in all of these results is based on using a cumulative GPA as the measure of ability. This raises the issue of whether alternative measures of ability may be as effective or more effective. As future cohorts complete this course it may be possible to combine their data with this cohort in order to use their entry level ranking as the proxy for underlying ability instead. It was not possible in this paper because not enough students had this data[8]. This could also provide information as to how good New Zealand school grades are at predicting success in a student's first year at university.

Not included in this analysis are student responses on the usefulness of this approach and of the separate components to their learning. Unfortunately it was never considered to have questions specific on the new assessment format inserted into the formal university course evaluations. However, the weekly quizzes were frequently referred to in the open-ended questions. There were two such questions on the evaluation form used. The first was 'What was most helpful to your learning?' In response to this question approximately 60% of the open-ended positive responses were on the quizzes, 30% on the tutorials and 10% miscellaneous other. The second question was 'What improvements would you like to see?' The same topics generally appeared in these responses as in the positive comments but weighted differently: 10% of the open-ended responses included ways to improve the quizzes, approximately 30% on the tutorials, 30% were about having access to old exam papers and 30% miscellaneous other topics including course content volume and pace, mark weightings, lecturers etc. Further commentary on these responses can be found in Hedges (Hedges, 2012).

Most of the improvement quiz problems were more about the organisation and management of the online platform. There were a number of issues with this early in the semester that were largely corrected later and will not be issues in future iterations of this course. The tutorial comments, both positive and negative were primarily related to specific tutors. It is expected,

[8] In the New Zealand system there are five different university entry qualifications that can be used. Only two of these generate an entry ranking. Two others are historical entry qualifications that are not directly transferrable to the current standards. Further details on these can be found in the University of Auckland Calendar (The University of Auckland, 2011).

and in fact desirable, that there will always be some variation between the tutors but over time it is hoped the degree and learning impact aspects of this variation can be minimised with better tutor training. The comments regarding access to old exam papers was an acknowledged by the staff, however because this was the first time the course had been offered in this form this was simply not possible. This had been explained to the students along with specific direction that their tutorial and test questions would provide the most useful guides for their exam preparation. They already had access to these and the marking guidelines that had been used. Clearly this issue will be solved with time.

Overall the course structure appears to have achieved it aims. Like any first iteration of a new course there is room for improvement in content, delivery and measurement but sufficient support was found for the overall structure and the quizzes in particular that this should be retained and opportunities for further analyses of these course design and assessment innovations.

References

Anderson, G., Benjamin, D., & Fuss, M. A. (1994). The Determinants of Success in University Introductory Economics Courses. *Journal of Economic Education, 25*(2), 99-119.

Ashford, J. R., & Sowden, R. R. (1970). Multivariate probit analysis. *Biometrics, 26*(3), 535-546.

Axelson, R. D., & Flick, A. (2010). Defining Student Engagement. *Change: Magazine of Higher Learning, 43*(1), 38-43.

Biggs, J. (1999). Teaching for Quality Learning at University. Buckingham: Society for Research into Higher Education/Open University Press.

Biggs, J. (2003). Aligning Teaching and Assessment to Curriculum Objectives (Imaginative Curriculum Project, Trans.). York: Learning and Teaching Support Network (Now Higher Academy).

Bomia, L., Beluzo, L., Demeester, D., Elander, K., Johnson, M., & Sheldon, B. (1997). The impact of teaching strategies on intrinsic motivation. *ERIC Clearinghouse on Elementary and Early Childhood Education*, 294. Retrieved from http://www.eric.ed.gov/PDFS/ED418925.pdf

Buckles, S., & Siegfried, J. J. (2006). Using Multiple-Choice Questions to Evaluate In-Depth Learning in Economics. *Journal of Economic Education, 37*(1), 48-57.

Cameron, M. (2010). *'Economics with Training Wheels': Using Weblogs in Teaching and Assessing Introductory Economics*. Paper presented at the XVth Australasian Teaching Economics Conference Hamilton, NZ.

Davies, P. (2003). Threshold Concepts: how can we recognise them? In J. H. F. Meyer & R. Land (Eds.), *Overcoming barriers to Student Understanding: Threshold concepts and troublesome knowledge* (pp. 70-84). London: Routledge.

Davies, P., & Guest, R. (2010). What effect do we really have on students' understanding and attitudes? How do we know? [Editorial]. *International Review of Economics Education, 9*(1).

Davies, P., & Mangan, J. (2005). *Recognising Threshold Concepts: an exploration of different approaches*. Paper presented at the European Association in Learning and Instruction Conference, Nicosia, Cyprus.

Davies, P., & Mangan, J. (2008). Embedding Threshold Concepts: from theory to pedagogical principles to learning activities. In R. Land, J. H. F. Meyer & J. Smith (Eds.), *Threshold Concepts within the Discipline* (pp. 37-50). Rotterdam: Sense Publishers.

Duff, A. (2004). Understanding academic performance and progression of first-year accounting and business economics undergraduates: the role of approaches to learning and prior academic achievement. *Accounting Education, 13*(4), 409-430. doi: 10.1080/0963928042000306800

Fizel, J. L., & Johnson, J. D. (1986). The Effect of Macro/Micro Course Sequencing on Learning and Attitudes in Principles of Economics. *Journal of Economics Education, 17*(2), 12.

Galizzi, M. (2010). An assessment of the impact of online quizzes and textbook resources on students' learning. *International Review of Economics Education, 9*(1), 31-43.

Gibbs, G. (1999). Using Asessment Strategically to Change the Way Students Learn. In S. Brown & A. Glasner (Eds.), *Assessment Matters in Higher Education* (pp. 41-53). Oxford, UK: SRHE/Oxford UP.

Gorinski, R., & Abernethy, G. (2007). Maori Student Retention and Success: Curriculum, Pedagogy and Relationships

Handbook of Teacher Education. In T. Townsend & R. Bates (Eds.), (pp. 229-240): Springer Netherlands.

Hedges, M. R. (2012). *Constructive ALignment: Linking to Business Beyond the Classroom*. Paper presented at the 17th Australasian Teaching Economics Conference, Gold Coast, Queensland, Australia.

Hickson, S. (2010, 30 June-2 July). *The Impact of Question Format in Principles of Economics Classes: Evidence from New Zealand.* Paper presented at the 52nd Annual Conference, New Zealand Association of Economists, Auckland, NZ.

Kuh, G. D. (2010). What We're Learning About Studnet Engagement From NSSE: Benchmarks for Effective Educaitonal Practices. *Change: Magazine of Higher Learning, 35*(2), 24-32.

Land, R., Cousin, G., Meyer, J. H. F., & Davies, P. (2005). Threshold concepts and troublesome knowledge: implications for course design and evaluation. In C. Rust (Ed.), *Improving Student Learning Diversity and Inclusivity*. Oxford: Oxford Centre for Staff Learning and Development.

Land, R., & Meyer, J. H. F. (Eds.). (2006). *Overcoming Barriers to Student Understanding.* London: Routledge.

Meyer, J. H. F., & Land, R. (2003). Threshold Concepts and Troublesome Knowledge (1): linkages to ways of thinking and practising witihin disciplines *Improving Student Learning - Ten Years On* (pp. 745-424). Oxford: Oxford Centre for Staff and Learning Development (OCSLD).

Meyer, J. H. F., & Land, R. (2005). Threshold Concepts and Troublesome Knowledge (2): Epistemological Considerations and a Conceptual Framework for Teaching and Learning. *Higher Education, 49*(3), 16.

Newmann, F. (1992). Student Engagement and Achievement in American Secondary Schools. *Teachers College Press.*

Park, K. H., & Kerr, P. M. (1990). Determinants of Academic Performance: A Multinomial Approach. *Journal of Economic Education, 21*(2), 101-111.

Rowntree, D. (1987). *Assessing Students: How shall we know them?* London: Kogan Page.

Schlechty, P. (1994). Increasing Student Engagement. *Missouri Leadership Academy*, 5.

Shanahan, M., Foster, G., & Meyer, J. H. F. (2008). Associations among prior acquisition of threshold concepts, learning dimesnions and examination performance in first-year economics. In R. Land, J. H. F. Meyer & J. Smith (Eds.), *Threshold Concepts within the Disciplines* (pp. 155-172). Rotterdam: Sense Publishers.

Staffordshire University. (2008). Embedding Threshold Concepts, from http://www.staffs.ac.uk/schools/business/iepr/etc/index.htm

Taylor, R. (2002). Designing Undergraduate Degree Programmes *Handbook for Economics Lecturers.* Bristol: Economics Network.

The University of Auckland. (2011). *The Univesrity of Auckland 2011 Calendar.* Auckland: The University of Auckland.

University of Auckland Business School (2009). [Undergraduate Curriculum Review 2009].

Zellner, A., & Lee, T. H. (1965). Joint estimation of relationships involving discrete random variables. *Econometrica, 33*(2), 382-394.

Table 1: Descriptive Statistics

Variables	Description	Mean (Stddev)
Final exam mark	Bounded variable: 0 – 50 (Final Exam accounted for 50% of assessment for the paper)	30.553 (9.005)
Male	Dummy variable: 1 = Male; 0 otherwise	0.557 (0.497)
Age	Age in years at the start of the semester	19.466 (3.479)
MaPP	Dummy variable: 1 = Maori or Pacific Peoples; 0 otherwise	0.101 (0.301)
Asian	Dummy variable: 1 = Asian; 0 otherwise	0.559 (0.497)
Ethnicities other than Maori, Pacific Peoples and Asian serve as the control group (this is predominantly Pakeha)		
Domestic	Dummy variable: 1 = NZ citizen or Permanent resident in NZ; 0 otherwise	0.786 (0.411)
Bbim	Dummy variable: 1 = Enrolled in Bachelor of Business Information Management degree; 0 otherwise	0.153 (0.360)
Other degree	Dummy variable: 1 = Enrolled in other degree; 0 otherwise (these are predominantly Bachelors of Arts or Property)	0.218 (0.413)
Enrolment in a Bachelor of Commerce (BCom) degree serves as the control group		
Ability	Continuous variable of student's cumulative GPA: 0 – 8.75 (0 equates to a D average; 9 equates to an A+ average)	3.575 (2.124)
Papers taken	Number of papers enrolled in semester: Continuous variable (2 – 5)	4.013 (0.303)
Quiz mark	Bounded variable: 0 – 10 (Quizzes accounted for 10% of assessment for the paper)	7.731 (2.139)
Total Time Spent on Quizzes	Number of hours spent on quiz attempts over the entire semester	11.868 (9.177)
Quiz efficacy	Quiz mark interacted with Total Time Spent on Quizzes	101.383 (91.150)
Tutorial mark	Bounded variable: 0 – 10 (Tutorials accounted for 10% of assessment for the paper)	8.109 (2.542)
Test mark	Bounded variable: 0 – 30 (The test accounted for 30% of assessment for the paper)	18.764 (4.941)

Figure 1: Tree diagram (Productive engagement based on quiz efficacy variable)

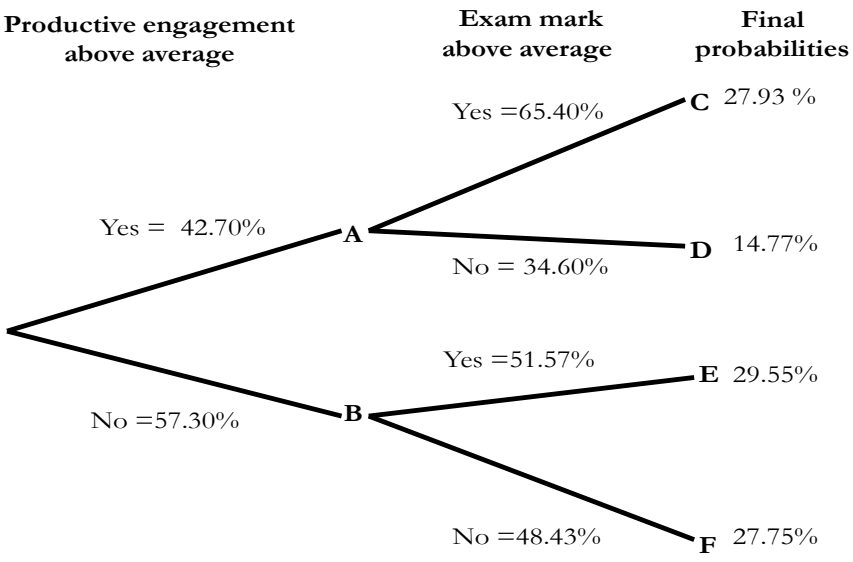

Table 2 – Linear regressions

Variables	(1)	(2)	(3)
Constant	19.475*** (5.634)	17.920*** (5.917)	8.387* (5.142)
Male	1.042** (0.499)	1.066** (0.497)	0.739 (0.496)
Age	-0.137 (0.153)	-0.137 (0.154)	-0.078 (0.135)
MaPP	-0.033 (1.005)	0.344 (0.984)	-0.182 (0.966)
Asian	-0.319 (0.555)	-0.343 (0.560)	-0.997* (0.558)
Domestic	-1.261* (0.682)	-1.269* (0.679)	-0.463 (0.631)
Bbim	-0.187 (0.768)	-0.093 (0.761)	0.018 (0.708)
Other degree	0.135 (0.585)	0.201 (0.579)	0.465 (0.569)
Ability	3.181*** (0.135)	3.051*** (0.135)	2.137*** (0.215)
Papers taken	0.571 (0.873)	0.533 (0.888)	0.971 (0.816)
Quiz efficacy	0.007*** (0.002)	0.005** (0.002)	0.005* (0.002)
Tutorial mark	-	0.274 (0.174)	0.290* (0.163)
Test mark	-	-	0.517*** (0.086)
R squared	0.586***	0.623***	0.625***

Notes: Robust standard errors shown in parenthesis. ***, ** and * represent statistical confidence at the 1%, 5% and 10% levels. Reference groups = Female, Pakeha, International students, and BCom degree. N = 555.

Table 3: Coefficient estimates in bivariate probit model

Variables	(1) Quiz efficacy	(2) Exam performance
Constant	-2.989*** (0.976)	-4.225*** (1.222)
Male	-0.353*** (0.115)	0.176 (0.144)
Age	0.003 (0.015)	0.003 (0.019)
MaPP	0.271 (0.238)	0.043 (0.262)
Asian	0.537*** (0.133)	0.090 (0.169)
Domestic	-0.063 (0.150)	-0.054 (0.194)
Bbim	-0.076 (0.161)	0.041 (0.180)
Other degree	-0.169 (0.153)	0.256 (0.196)
Ability	0.064** (0.032)	0.590*** (0.070)
Papers taken	0.181 (0.205)	0.223 (0.237)
Tutorial mark	0.205*** (0.035)	-0.023 (0.035)
Test mark	-	0.085*** (0.021)
N	555	
Log pseudo likelihood	-525.983	
Rho	-0.052 (0.095)	

Notes: Robust standard errors shown in parenthesis. ***, ** and * represent statistical confidence at the 1%, 5% and 10% levels. Rho suggests positive correlation between regressions ($chi^2(1)=0.300$, $p<0.000$).

Table 4: Marginal effects

Variables	(1) Exam performance given Quiz efficacy = 1	(2) Exam performance given Quiz efficacy = 0
Male	0.062 (0.055)	0..060 (0.053)
Age	0.001 (0.007)	0.001 (0.007)
MaPP	0.019 (0.097)	0.018 (0.094)
Asian	0.041 (0.065)	0.039 (0.062)
Domestic	-0.021 (0.072)	-0.020 (0.069)
Bbim	0.014 (0.067)	0.014 (0.065)
Other degree	0.091 (0.069)	0.088 (0.066)
Ability	0.222***(0.027)	0.214*** (0.023)
Papers taken	0.086 (0.090)	0.083 (0.086)
Tutorial mark	-0.006 (0.014)	-0.006 (0.013)
Test mark	0.032*** (0.008)	0.031*** (0.008)

Notes: Robust standard errors shown in parenthesis. ***, ** and * represent statistical confidence at the 1%, 5% and 10% levels.